DOCUMENT RESUME

ED 328 578 TM 016 085

AUTHOR Fisk, Yvette Hester
TITLE A Brief Overview of Three Classes of Methods for Detecting Item Bias.
PUB DATE Jan 91
NOTE 26p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (San Antonio, TX, January 24-25, 1991).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Chi Square; Comparative Analysis; *Difficulty Level; *Item Bias; Item Response Theory; Literature Reviews; Mathematical Models; *Statistical Analysis; Test Items
IDENTIFIERS Delta Plot Method

ABSTRACT

The reasons for recent endeavors to evaluate item bias are discussed, and item bias is defined. Some of the literature regarding the most promising methods of detecting item bias is reviewed. Three classes of methods for detecting item bias are discussed using concrete examples and illustrations. These methods are: (1) latent trait; (2) chi-square; and (3) item difficulty methods. The item difficulty method is considered to be the least computationally demanding method of the three. A small data set consisting of two groups of 17 examinees each taking a 15-item dichotomously-scored test is provided to illustrate the delta-plot technique--a user-friendly method involving item difficulties. It is concluded that researchers must decide on the best method to use based on: (1) the application of the results; (2) the level of statistical sophistication required; and (3) the practicality of implementing each method. Latent trait methods are statistically superior to chi square and item difficulty methods; however, the latter two methods are easier to implement. A 28-item list of references, two tables, and four graphs are included. (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED328578

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Yvette Hester Fisk

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

A BRIEF OVERVIEW OF THREE CLASSES OF METHODS FOR DETECTING ITEM BIAS

Yvette Hester Fisk

Texas A&M University

Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, January 25, 1991.

TM016085

ABSTRACT

The reasons for recent endeavors to evaluate item bias are discussed and item bias is defined. Some of the literature regarding the most promising methods of detecting item bias is reviewed. Three classes of methods for detecting item bias are discussed using concrete examples and illustrations. These three classes of methods are latent trait, chi-square, and item difficulty methods. The item difficulty method is elaborated in some detail, since this method is computationally least demanding. A small data set is provided to illustrate the delta-plot technique, a user-friendly method involving item difficulties.

The word "bias" carries certain unfavorable connotations regardless of the context in which it is used. There are social implications connected with this word and it has a highly disputatious nature. Thus, the measurement community decided to work toward developing a body of rigorous, empirical research during the late sixites and early seventies to elaborate the construct of bias. The mandate for this work was also articulated in the major legal cases on bias, e.g., *Larry P., et al.* v. *Wilson Riles, et al.*, 1972, 1979 (Berk, 1982).

Bias is not easy to evaluate in reference to any measurement procedure or tool. For example, saying a test is unbiased test does not imply that all the specific items that make up the test are unbiased. Therefore, item bias studies are necessary for evaluating all tests and should be conducted early in the stages of test construction to reduce possible bias. Although, the elimination of item bias may reduce the probability of test bias, additional research is needed to verify that this latter objective has been realized (Berk, 1982).

The interaction that is of specific interest in an item bias study is the item-by-group interaction. A biased item favors or disfavors one group over another, a situation both socially and psychometrically undesirable. Every attempt should be made to eliminate or, at the very least, minimize bias (Jensen, 1980). According to Berk (1982), the study or examination of item bias should include: (a) a judgmental review involving a panel of minority group judges; (b) statistical analysis; and (c) a posteriori statistical analysis to determine whether bias was present, why it was present and to consider procedures for eliminating the bias. Berk (1982, p. 4) notes, however, that "While the research comparing the data from judgmental and statistical

methods has found almost irreconciliable differences between the methods (see Cotter & Berk, 1981; Plake, 1980; Sandoval & Miille, 1980), a more perceptive interpretation might be that the two methods are tapping different kinds of bias."

Reynolds (1982) also questions the exclusive effectiveness of expert panels based on the dismal evidence for the method's effectiveness. Perhaps the best role for the "logical analyses panel" is determining the correct usages of the test, and detecting any offensive formats, such as stereotypical protrayals of specific groups, that may be ignored or overlooked by the empirical methods (Shepard, 1982). One must bear in mind that the many statisitical methods included in the present literature on item bias are meant only as extentions or supplements to the test developers' judgement, not as substitutes (Angoff, 1982).

According to Shepard (1982, p. 21), "unbiasedness in test items is best conceptualized as equal construct validity across groups and must be confirmed by verifying expected patterns of relationships." With this conceptual definition in mind, note that items typically are written with the intention of measuring the construct of interest. Therefore, reasonably appropriate content should be examined for possible sources of variation that are irrelevant or in which the researcher is not interested. These irrelevant sources of variation are often overlooked by a subjective examination, since they are usually not obvious. Thus, methods for an empirical analysis are used. These are statistical techniques that search out items that do not contribute the same meaning to the total test for each group. Thus, formal research definitions for an unbiased item include:

> A test item can be considered to be unbiased if all individuals having the same underlying ability level have an equal probability of correctly answering the item, regardless of their subgroup membership. (Pine, 1977, p. 38)

> An item is unbiased, if, for all individuals having the same score on a homogeneous subtest containing the item, the proportion of individuals getting the item correct is the same for each population group being considered. (Scheuneman, 1975, 2).

> A set of items is unbiased if (1) the items are affected by the same source of variance in both subpopulations; and (2) among examinees who are at the same level on the construct purportedly measured by the test, the distributions of irrelevant sources of variation are the same for both subpopulations. (Crocker & Algina, 1986, p. 377)

The purpose of the present paper is to briefly examine the three most promising classes of methods for detecting item bias: latent trait, chi-square and item difficulty methods. A brief outline and discussion of each of these three classes of methods is presented. Methodological and procedural details have been omitted in part, but references have been provided for those who wish to pursue more indepth treatments. The item difficulty method is elaborated in more detail, since this method is computationally least demanding.

Latent Trait Methods

The first class of methods to be discussed are methods based on item response theory or latent trait theory. Lawson (1991) and McKinley and Mills (1989) provide

readable treatments of these sophisticated methods. These methods use item characteristic curves (ICCs) to portray the relationship between ability and the probability of answering a given item correctly. An item is said to be unbiased if the ICCs for every item are the same for all groups. Or conversely, biased items have different ICCs.

ICCs are given by cumulative logistic functions or normal ogives that are described using parameters. Some of the parameters which may be included are: (a) an item difficulty parameter (the ability level corresponding to the inflection point of the curve); (b) an item discrimination value (proportional to the slope of the curve at the inflection point); and (c) a guessing parameter (the probability of examinees of infinitely low ability getting the item correct) (Ironson, 1982).

Most of the indices of bias based on Item Response Theory are measures of the extent to which the ICCs vary across groups. Estimates of the item parameters for each group are calculated and expressed on the same scale and the ICCs are compared. An index of item bias is then calculated for each item. The particluar steps involved in carrying out these procedures is dependent on the model chosen. The latent trait models discussed here will be limited to the model using all three of the above parameters (the three-parameter model) and the one-parameter or Rasch model.

The three-parameter model is the most theoretically sound and statistically complex procedure for measuring bias (Bejar, 1980; Hunter, 1975; Lord, 1980; Peterson, 1977). This model requires a high degree of statistical sophistication, a program (LOGIST or a similar program) that is costly to run, and a large sample size

(Ironson, 1982). The measures most often used to index bias for this three-parameter model are: (a) the area between the curve; and (b) the hypothesis test of the equality of the three parameters across ethnic groups (Lord, 1977, 1980). Below are some examples of possible ICCs obtained by three-parameter models.

INSERT FIGURE 1 ABOUT HERE

Figure 1 illustrates ICCs for an item that is biased against a specific group. In general, for equal ability, the group whose ICC is above the other has a greater chance of getting the item correct. Thus, in figure 1, since the ICC for males lies consistently below the ICC for females, the item is biased against males.

INSERT FIGURE 2 ABOUT HERE

Figure 2 illustrates nonuniform bias. At the low ability range, the item is biased against females, since the ICC for females lies below the ICC for males. At the high ability range, the item is biased against males, since the ICC for males lies below the ICC for females.

An estimate of bias is given by the area between the ICCs. To calculate the area, one typically uses a Reimann sum approximation. That is, if the probabilities $P_{ig}(q)$ are given on some interval $a \leq q \leq b$ in increments of $(b - a)/n$ for some positive integer n, then the Reimann sum yielding the approximate area A_g is given by

$$A_g = \sum_{k=1}^{n} \frac{(b-a)}{n} \left| P_{1g}\left(a+k\frac{(b-a)}{n}\right) - P_{2g}\left(a+k\frac{(b-a)}{n}\right) \right|$$

In other words, one partitions the figure into pieces, calculates areas for each piece, and then sums them. For example, Rudner (1977) employed this technique on the interval $-4.00 \leq q \leq 4.00$ in increments of .005. As noted by Crocker and Algina (1986), the absolute value signs in the formulation of $\mathbf{A}_g$ make all differences between the ICCs positive, so that the positive and negative differences cannot balance out. This is important since, if the differences balanced out exactly, $\mathbf{A}_g = 0$ would be a possible result for a biased item, which would be misleading.

Lord (1980) proposed an alternative hypothesis test procedure where the null hypothesis is given by:

$$\mathbf{H_o}: \; a_{1g} = a_{2g}, \quad b_{1g} = b_{2g}.$$

This is used in conjunction with the estimation procedure that standardizes on the b_g's for the groups. The guessing parameter is difficult to estimate accurately and is, therefore, not included in the test. The statistic used is chi-square, which is an estimate of the degree of bias in the item. If the null hypothesis is rejected for an item, the item may be considered biased. A similar test is available for a two-parameter model. (Crocker & Algina, 1986)

The one-parameter model or Rasch model is based on the work of Georg Rasch (1960, 1966, 1980) and uses the item difficulty parameter as the sole estimate of performance. The procedure sets the discrimination of the items to a constant which implies that all items are equally discriminating. This model is less complex than the three-parameter model, requires the program BICAL (or a similar program), which is less costly to run than LOGIST, and requires a comparatively smaller sample size.

The one-parameter model is sometimes viewed as a special case of the three-parameter model in which the only parameter that is allowed to vary is the difficulty parameter.

A common procedure for this method is to examine the differences in difficulty levels of an item across ethnic groups. Again, a chi-square statistic is computed for each item in each group, to help determine fit (Ironson, 1982). Figure 3 illustrates a possible ICC obtained by using this procedure for a one-parameter or Rasch model.

INSERT FIGURE 3 ABOUT HERE

This ICC is typical of an item that is biased against one ethnic group over another. Note that the ICC for whites is consistently higher than the ICC for blacks. This indicates that the item is biased against blacks. Also note that the bias is greatest for the middle ability range, since the distance between the two curves is the greatest over the middle interval. For extremes, i.e., the lowest ability levels and the highest ability levels, the bias is almost zero.

For the class of methods based on latent trait theory, if the ICCs are equal, then the items are unbiased. It should be noted, however, that isolating ICCs for an item that differs across groups may indicate multidimensionality and not bias, if a unidimensional model, such as the Rasch model, has been used. The disadvantage is that dimensionality of items in each group must be established and there is currently no entirely adequate procedure for this process (Crocker & Algina, 1986).

Chi-square Techniques

Chi-square techniques "essentially define an item as biased if, within a group of examinees with scores in the same test score interval, the proportion of examinees responding correctly to the item is the same for both subpopulations." (Crocker & Algina, 1986, p. 383) These techniques are considered approximations to item response theory techniques, but are easier to employ. Note that the observed test score has replaced latent trait in ability estimates analyses to detect a biased item.

Two chi-square techniques are briefly discussed in this section. One technique is $\chi^2_{correct}$ (Scheuneman, 1975, 1979) and the other technique is χ^2_{full} (Camilli, 1979; Nungester, 1977). Scheuneman (1975, p. 2) focuses on correct responses to a given item and gives the following definition: "an item is unbiased if, for all individuals having the same score on a homogeneous subtest containing the item, the proportion of individuals getting the item correct is the same for each population group being considered." Camilli (1979) focuses on both correct and incorrect responses to a given item. An outline of the procedure is as follows (Crocker & Algina, 1986; Ironson, 1982):

1. Divide the observed score scale into several ability or score levels.
    A. A minimum of 10 to 20 observed correct responses should be included in each interval.
    B. Expected frequencies should be at least 5 per interval and a minimum number of incorrect responses to be included in each interval should be decided upon.

C. If using χ^2_{full}, the same guidelines for incorrect responses and expected frequencies should be used as was used for correct responses.

D. Three to five intervals are typically used. Cutoff scores for these intervals is a critical issue as changes here could easily alter the magnitude of the chi-square.

2. Compute χ^2.

3. Test the statistic for significance.

   A. Within each interval, groups are compared with respect to the proportion responding correctly to an item. The item is biased if the proportions vary across groups.

   B. Statistical significance implies that the item is biased.

   C. The magnitude of the statistic may be viewed as an index of the amount of bias.

The two methods may require different cutoff scores for the intervals used. Also, the sample size required for $\chi^2_{correct}$ is smaller. The two methods are in agreement mostly on items of higher difficulty and are least in agreement on items of low difficulty. Both methods are more conceptually straightforward and both require smaller sample sizes than item response theory techniques. (Ironson, 1982) A disadvantage with chi-square techniques in general is that measurement error could be reflected as item bias. Unreliability of the data may be detected by this technique as item bias.

As noted by Crocker and Algina (1986, p. 387) correlational studies (Ironson & Subkoviak, 1979; Rudner, Getson & Knight, 1980) suggest that "when the items are unidimensional in both groups and when item bias exists" $\chi^2_{correct}$ "will be almost as accurate an indicator of the amount of bias as $\mathbf{A}_g$ will be." Ironson (1982, pp.153-154) gives a thorough summary of advantages and disadvantages of the two chi-square techniques previously discussed, along with a summary of latent trait technique advantages and disadvantages.

Item Difficulty Techniques

This section reviews the delta-plot method for detecting item bias, which is based on item difficulty values (p) and was developed by Angoff (1972; also Angoff & Ford, 1973). For this method, a set of items is unbiased if the item difficulties for each group are perfectly correlated. Item difficulty values for two different groups are computed on a number of items that sample a single content or skill domain. Then the p-values are converted to a normal deviate with an arbitrary mean and standard deviation. In this case, and typically, the values 13 and 4 are used, respectively. These transformed normal deviates are called *deltas* and are plotted as ordered pairs on a bivariate graph. The delta measure of item difficulty is given by:

$$\Delta_g = 4z_g + 13,$$

where z_g is the cutoff score for the p-value of the gth item.

A high correlation, e.g., .98 is common if the rank order of difficulty is essentially the same for both groups. Thus, it is the outliers that contribute the most to the

item-by-group interaction, which is the interaction of most interest to the researcher studying item bias. The outliers imply these items have a different "psychological meaning" for members of each group (Angoff, 1982). Outliers are considered possibly biased items.

EXAMPLE OF DELTA METHOD

Assume the data in Table 1 were collected on two groups of 17 examinees each, on a 15 item, dichotomously scored test.

INSERT TABLE 1 ABOUT HERE

The item difficulty value P_{ig} for the i^{th} item and the g^{th} group is computed by:

P_{ig} = (total # correct)/(total # of examinees).

The z-score cut offs for each proportion in the standard normal distribution is listed for each group and the delta values are calculated as $\Delta = 4\ z + 13$. Each of these calculated values is presented in Table 2.

INSERT TABLE 2 ABOUT HERE

The delta plot with the delta values for group1 on the x-axis and group 2 on the y-axis is presented in Figure 4.

INSERT FIGURE 4 ABOUT HERE

The two outliers (items 2 and 11), indicated by the symbol +, would be considered possibly biased items, since they appear to lie off of the line of best fit.

A formal way to measure the distance of each outlier from the major axis of the ellipse formed by the scatter plot of the data is given by Angoff (1982). The departure (d) from the line $y = ax + b$ is calculated by:

$$d_i = \frac{ax_i - y_i + b}{\sqrt{a^2 + 1}}$$

where x and y are the delta values for the two groups on the ith item and the slope (a) is given by:

$$a = \frac{(s_y^2 - s_x^2) \pm \sqrt{(s_y^2 - s_x^2) + 4r_{xy}^2 s_x^2 s_y^2}}{2r_{xy} s_x s_y}$$

The standard deviation of the two groups and the correlation between the delta values are needed to calculate this slope. The y-intercept (b) is given by:

$$b = M_y - aM_x,$$

where M is the mean of the deltas for the particular group.

One of the major drawbacks of the delta-plot method is the fact that it is greatly influenced by the distribution of the ability of the examinees (Ironson, 1982). Angoff (1982, p. 106) lists a complete chart of the advantages and disadvantages of this method and concludes that, "The criticisms of the delta-plot procedure are quite valid. With appropriate controls for differences in the abilities of the major and minor groups and with an adjustment for item-test correlation, however, it is likely that the method

will yield results more closely approximating those given by the more sensitive item characteristic curve (ICC) methods. In addition, the delta-plot procedure is clearly far simpler and much less expensive than ICC methods, and it has applications in some types of cross-cultural studies not shared by the other methods". A personal computer and spread-sheet software lend themselves well to this method.

Summary

Some literature regarding the three most promising classes of methods for detecting item bias was reviewed. These three methods are latent trait, chi-square, and item difficulty methods. The researcher must decide on the best method to use based on the application of the results, the level of statistical sophistication required, and the practicality of implementing each method. The latent trait methods are statistically superior to chi-square and item difficulty methods. The latter are easier to implement. A small data set was employed to illustrate the computational ease of the delta-plot method; a method involving item difficulties.

References

Angoff, W.H. (1982) Uses of difficulty and discrimination indices for detecting item bias. In R.A. Berk (Ed.). *Handbook of methods for detecting test bias,* pp. 96-116. Baltimore: Johns Hopkins University Press.

Bejar, I.I. Biased assessment of program impact due to psychometric artifacts. *Psychological Bulletin,* 1980, *87,* 513-524.

Berk, R.A. (1982) Introduction. In R.A. Berk (Ed.). *Handbook of methods for detecting test bias,* pp. 1-8. Baltimore: Johns Hopkins University Press.

Camilli, G. (1979). A critque of the chi-square method for assessing item bias. Unpublished paper, Laboratory of Educational Research, University of Colorado.

Cotter, D.E. & Berk, R.A. (April, 1981) *Item bias in the WISC-R using black, white, and Hispanic learning disabled children.* Paper presented at the annual meeting of the American Educational Research Association, Los Angeles.

Crocker, L. & Algina, J. (1986) Detecting item bias. *Introduction to classical and modern test theory,* pp. 376-395. Holt, Rinehart and Winston, Inc.

Hunter, J.E. (December, 1975) *A critical analysis of the use of item means and item-test correlations to determine th presence or absence of content bias in achievement test items.* Paper presented at the National Institute of Education Conference on Test Bias, Annapolis, Md.

Ironson, G.H. (1982). Use of chi-square and latent trait approaches for detecting item bias. In R.A. Berk (Ed.). *Handbook of methods for detecting test bias,* pp. 117-160. Baltimore: Johns Hopkins University Press.

Ironson, G.H., & Subkoviak, M. A comparison of several methods of assessing item

bias. *Journal of Educational Measurement*, 1979, *16*, 209-225.

Jensen, A.R. (1980) *Bias in mental testing.* New York: Free Press.

Lawson, S. (1991). One parameter latent trait measurement: Do the results justify the effort? In B. Thompson (Ed.), *Advances in educational research: Substantive findings, methodological developments* (Vol. 1, pp. 159-168). Greenwich, CT: JAI Press.

Lord, F.M. A study of item bias, using item characteristic curve theory. In Y.H. Poortinga (Ed.), *Basic problems in cross-cultural psychology.* Amsterdam: Swets and Zeitlinger, 1977. Pp. 19-29.

Lord, F.M. *Applications of item response theory to practical testing problems.* Hillsdale, N.J.: Lawrence Erlbaum Associates, 1980.

McKinley, R.L., & Mill, C.N. (1989). Item response theory: Advances in achievement and attitude measurement. In B. Thompson (Ed.), *Advances in social science methodology* (vol. 1, pp. 71-135). Greenwich, CT: JAI Press.

Nungester, R.J. An empirical examination of three models of item bias. (Doctoral dissertation, Florida State University, 1977). *Dissertation Abstracts International*, 1977, *38*, 2726A. (University Microfilms No. 77-24, 289)

Petersen, N.S. *Bias in the the selection rule: Bias in the test.* Paper presented at the Third International Symposium on Educational Testing, University of Leyden, Netherlands, June 1977.

Pine, S.M. Applications of item characteristic curve theory to the problem of test bias. In D.J. Weiss (Ed.), *Applications of computerized adaptive testing* (RR 77-1). Minneapolis: Department of Psychology, Psychometric Methods Program,

University of Minnesota, March, 1977. Pp. 37-43.

Plake, B.S. A comparison of a statistical and subjective procedure to ascertain item validity: One step in the test validation process. *Educational and Psychological Measurement,* 1980, *40*, 397-404.

Rasch, G. *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Denmarks Paedagogiske Institut, 1960.

Rasch, G. An individualistic approach to item analysis. In P.F. Lazarsfeld & N.W. Henry (Eds.), *Readings in mathematical social science.* Chicago: Science Research Associates, 1966. Pp.89-108.

Rasch, G. *Probabilistic models for some intelligence and attainment tests.* Chicago: University of Chicago Press, 1980.

Reynolds, C.R. The problem of bias in psychological assessment. In C.R. Reynolds & T.B. Gutkin (Eds.), *The handbook of school psychology.* New York: Wiley, 1982. Pp. 178-208.

Rudner, L.M. (1977). An approach to biased item identification using latent trait measurement theory. Paper presented at the annual meeting of the American Educational Research Association, New York.

Rudner, L.M., Getson, P.R., & Knight, D.L. A Monte Carlo comparison of seven biased item detection techniques. *Journal of Educational Measurement,* 1980, *17*, 1-10.

Sandoval, J., & Miille, M.P.W. Accuracy of judgments of WISC-R item difficulty for minority groups. *Journal of Consulting and Clinical Psychology,* 1980, *48*, 249-253.

Scheuneman, J.D. *A new method of assessing bias in test items.* Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C., April 1975. (ERIC Document Reproduction Service No. ED 106 359)

Scheuneman, J.D. A new method for assessing bias in test items. *Journal of Educational Measurement*, 1979, *16*, 143-152.

Shepard, L.A. Definitions of Bias. In R.A. Berk (Ed.). *Handbook of methods for detecting test bias*, pp. 9-30. Baltimore: Johns Hopkins University Press.

Table 1

Hypothetical Data Set for Delta-plot method

item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Examinee															
Group 1															
1	1	1	0	0	1	0	0	1	1	1	0	0	1	1	0
2	0	1	0	1	0	0	0	0	1	0	0	1	1	0	0
3	0	1	0	1	0	0	0	0	1	0	1	0	1	1	0
4	1	1	1	1	0	0	0	0	1	0	0	1	1	0	0
5	1	1	0	1	0	0	1	1	1	0	0	0	1	1	0
6	1	0	0	1	0	0	0	0	1	0	0	1	1	0	0
7	0	1	1	1	1	0	0	0	1	0	1	0	1	1	0
8	1	0	1	1	1	1	0	0	1	0	0	1	1	0	0
9	0	1	0	1	1	0	1	1	1	0	0	0	0	0	0
10	0	1	0	1	1	0	0	0	1	0	1	1	0	0	0
11	0	1	1	1	1	0	0	0	1	0	1	1	0	1	1
12	0	1	0	1	0	0	0	0	1	0	0	1	0	0	0
13	0	0	0	1	0	0	0	1	1	0	0	1	0	0	0
14	0	1	0	1	0	0	0	0	1	0	1	1	0	1	0
15	1	1	1	0	1	0	1	0	1	0	1	1	0	0	0
16	1	1	0	0	1	0	0	0	0	0	0	1	0	0	0
17	0	1	1	0	1	0	0	0	1	0	0	0	0	1	0
Group 2															
1	0	1	1	0	0	0	1	1	1	0	1	0	0	1	0
2	1	0	1	1	0	0	0	0	1	0	1	1	0	1	0
3	0	1	0	1	1	0	0	0	1	0	1	0	0	1	0
4	1	1	0	1	1	0	0	0	1	0	1	1	1	1	0
5	0	1	0	1	0	0	0	0	1	0	0	1	0	1	0
6	1	1	1	1	1	0	0	0	1	0	1	1	1	1	0
7	0	0	1	1	1	0	0	0	1	0	1	1	0	0	0
8	1	0	0	1	1	0	1	0	1	0	1	1	1	0	0
9	0	1	0	1	0	0	0	1	0	0	0	1	1	0	0
10	1	1	0	1	0	0	0	0	1	0	1	0	0	0	0
11	0	1	0	0	1	0	0	0	1	0	1	0	1	0	0
12	1	0	1	0	1	0	0	0	1	1	0	0	0	0	0
13	0	0	1	0	1	0	1	0	1	0	0	0	1	0	1
14	1	0	0	1	0	0	0	0	1	0	0	0	1	0	0
15	0	1	0	1	0	1	0	0	1	0	1	1	1	0	0
16	1	1	0	1	1	0	0	0	1	0	1	1	0	0	0
17	0	1	0	1	0	0	0	1	1	0	1	1	0	0	0

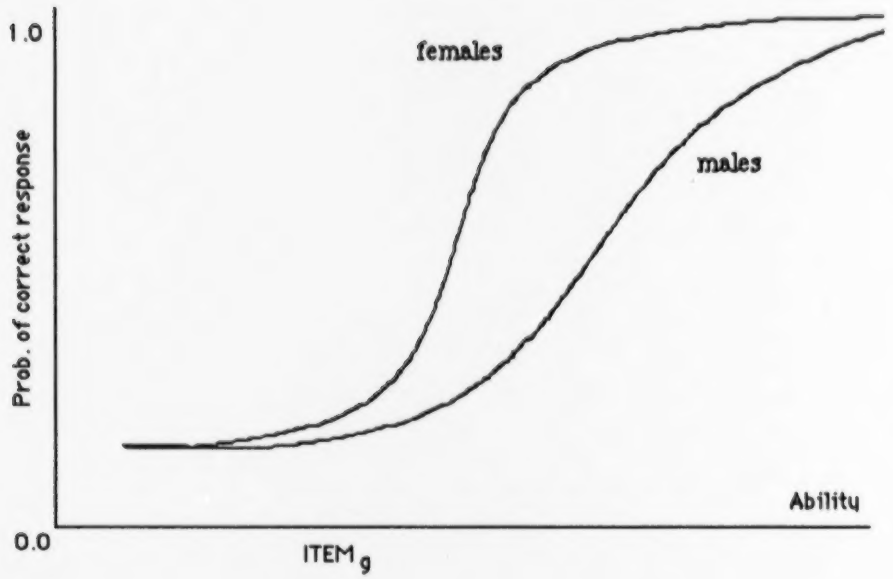
TABLE 2

Spread Sheet Calculations using Excel

Item	P (Group 1)	P (Group 2)	Z score (G1)	Z score (g2)	delta (g1)	delta (g2)
1	0.4118	0.4706	-0.2230	-0.0740	12.1080	12.7040
2	0.8235	0.6471	0.9305	0.3780	16.7220	14.5120
3	0.3529	0.3529	-0.3790	-0.3790	11.4840	11.4840
4	0.7647	0.7647	0.7210	0.7210	15.8840	15.8840
5	0.5294	0.5294	0.0740	0.0740	13.2960	13.2960
6	0.0588	0.0588	-1.5670	-1.5670	6.7320	6.7320
7	0.1765	0.1765	-0.9290	-0.9290	9.2840	9.2840
8	0.2353	0.1765	-0.7210	-0.9290	10.1160	9.2840
9	0.9412	0.9412	1.5650	1.5650	19.2600	19.2600
10	0.0588	0.0588	-1.5660	-1.5660	6.7360	6.7360
11	0.3529	0.7059	-0.3770	0.5410	11.4920	15.1640
12	0.6471	0.5882	0.3780	0.2220	14.5120	13.8880
13	0.4706	0.4706	-0.0740	-0.0740	12.7040	12.7040
14	0.4118	0.3529	-0.2230	-0.3770	12.1080	11.4920
15	0.0588	0.0588	-1.5660	-1.5660	6.7360	6.7360

Figure 1

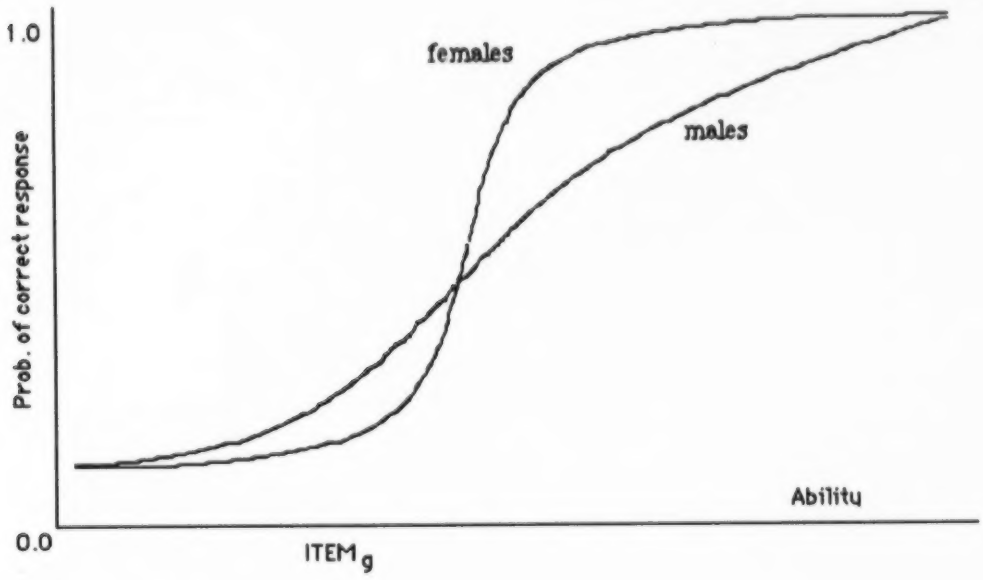
Hypothetical ICCs for a 3-paramenter model depicting item bias



($P_g(\theta)$ = probability of a correct response on item g as a function of ability)

Figure 2

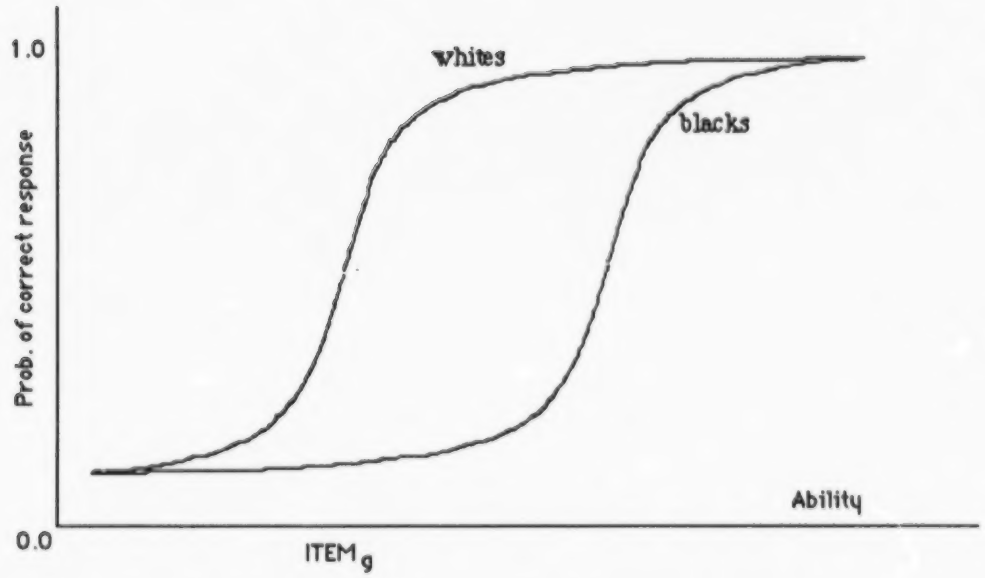
Hypothetical ICCs for 3-parameter model depicting nonuniform item bias



($P_g(\theta)$ = probability of a correct response on item g as a function of ability)

Figure 3

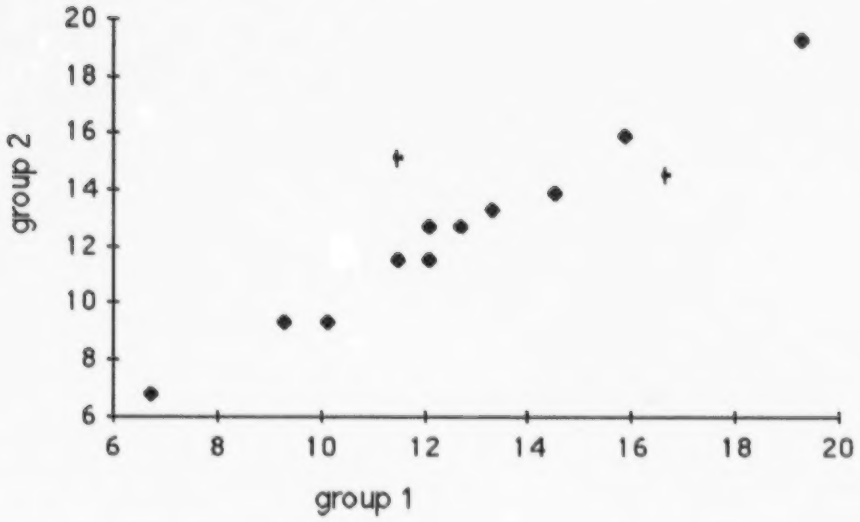
Hypothetical ICCs for a 1-parameter model depicting biased item



($P_g(\theta)$ = probability of a correct response on item g as a function of ability)

Figure 4

Delta-Plot



+ denotes possibly biased item